AD-A264 353

ENTATION PAGE

		FINAL/15 MAR 91 TO 14 SEP 92

OUTLIER DETECTION IN INFRARED SIGNATURES (u)

Dr. Michael Chernick

2304/A5
F49620-91-C-0029

Nichols Research Corp.
Newport Beach CA 92660

20

AFOSR/NM
110 DUNCAN AVE, SUTE B115
BOLLING AFB DC 20332-0001

F49620-91-C-0029

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

UL

For a number of years, simulated long wavelength infrared (LWIR) signatures have been used to determine the ability to classify military targets and decoys. Such signatures sometimes exhibit specular behavior, a characteristic displaying a sudden increase in radiant intensity of short duration. This specular behavior is sporadic and is as likely to show up for targets as it is for decoys. Unfortunately, if these outliers (i.e. the specular occurrances) are not removed from the data, the estimated performance of discrimination algorithms can be misleading. Statistical outlier detection provides a useful approach for finding and removing the outliers caused by specular occurrences.

93 5 1? 094

93-10737

23

SECURITY CLASSIFICATION OF REPORT	SECURITY CLASSIFICATION OF THIS PAGE	SECURITY CLASSIFICATION OF ABSTRACT	LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR(SAME AS REPORT)

Outlier Detection in Infrared Signatures

January 1992

Jon A. Magnuson
Nichols Research Corporation
Newport Beach, CA 92660

Mike R. Chernick
Risk Data Corporation
Irvine, CA, 92718

Abstract

For a number of years, simulated long wavelength infrared (LWIR) signatures have been used to determine the ability to classify military targets and decoys. Such signatures sometimes exhibit specular behavior, a characteristic displaying a sudden increase in radiant intensity of short duration. This specular behavior is sporadic and is as likely to show up for targets as it is for decoys. Unfortunately, if these outliers (i.e. the specular occurrences) are not removed from the data, the estimated performance of discrimination algorithms can be misleading. Statistical outlier detection provides an useful approach for finding and removing the outliers caused by specular occurrences.

This paper considers the statistical properties of the outlier detection algorithms as applied to simulated LWIR signatures. We consider possible statistical models for outliers in order to determine whether or not modifications might minimize the number of outliers left in the signature after editing and minimize the number of "good" observations deleted from the signature. Ultimately, we are seeking the data editing algorithm which produces the best possible discrimination performance.

1. Introduction

A small number of outliers in a data set can have a large influence on analyses done with the data. This is particularly true when the sample size is small and the number of variables is relatively large, as is often the case in real world problems. In such multivariate situations, outliers may not be easy to detect or define (i.e., What distance measure should be used to define an extreme observation?). Multiple outliers tend to mask each other making detection more difficult. When standard regression analyses are run, the presence of an outlier may not be apparent in the residuals even though their presence has had a large effect on the values of the regression parameters. In recent years it has been shown that even a single outlier can affect the collinearity structure of the regression variables (e.g. see Chatterjee and Hadi (1988) and Walker (1989)).

This work was supported in part by contract number 90NM110 with the Air Force Office of Scientific Research, Bolling Air Force Base.

In time series analysis it has been shown (see Chernick et. al (1982) for example) that outliers can make it appear, based on white noise tests of residuals, that a series is merely a white noise process when in fact it is really a contaminated first order autoregressive process. This points to a general problem, namely that outliers can hide the correlation structure in a time series. This correlation structure is crucial for model identification.

Gnanadesikan (1977) pointed out that Hampel's influence function (Hampel (1974)) can be used to estimate the effect individual outliers have on sample estimates of parameters. Chernick noted that the influence function for parameters of interest to the users of a data base provides a way of defining which outliers are important (i.e., observations with large estimated influence on parameters of interest to users are important outliers, while those with small estimated influence are not). In this way the influence function provides a "distance" measure for multi-variate outliers. This approach was applied by Chernick to the problem of data validation for the Department of Energy data bases in Chernick et al. (1982) and Chernick (1982a) and was continued as an approach to data editing in Chernick and Murthy (1983).

Recently some important practical work has been done by Lefrancois (1991) following up on the use of influence function measures in time series, first proposed in Chernick et al. (1982). His methods are considered in this paper. Earlier work following up on Chernick et al. (1982) are Lattin (1983) and Li and Hui (1987)

2. Detecting Specular Type Outliers in Signature Data

Infrared (IR) measurements are used by the military to help identify unknown targets that may have been sent by the enemy. Specific features of each IR time-series help to discriminate and classify the unknown targets automatically. Outliers in the time-series corrupt the estimation of these features and, if the corruption is bad enough, will cause the discrimination algorithm to reach wrong, or dangerous, conclusions. A computer outlier detection algorithm is necessary since the signals are coming in "real-time". Even though a human may be able to do a better job of outlier detection, he would be too slow to sort out the large number of possible time-series soon enough to use the information.

At Nichols Research Corporation there has been a significant amount of work done in recent years evaluating the performance of classification algorithms for the Strategic Defense Initiative (SDI) targets using simulated long wavelength infrared (LWIR) and laser radar (ladar) signatures. Specular behavior in the LWIR signature (a particular type of outlier in signatures), which is characterized by sudden increase in radiant intensity of short duration, occurs infrequently and sporadically in the simulations.

It is somewhat controversial whether or not these phenomena are artifacts of the Optical Signatures Code (the standard code used to simulate LWIR signatures) or represent real effects that an LWIR sensor would observe. In either case, due to their sporadic occurrence, these outliers could not be expected to provide useful information for discrimination and hence editing procedures for the detection and removal of specular observations had been routinely included in discrimination studies.

When an IR sensor is based in space, looking at a target in space, there is very little background radiation to corrupt the IR measurements from the target. Changes in IR intensity from the target can aid in describing the target, (see Figure 1). Many different features of IR signatures have been proposed to help discriminate and classify targets. The underlying physical parameters tend to be things like thermal mass and shape. There are more proposed features than can be adequately covered in a paper on outliers. All of the

A-1

features have the common problem that outliers will corrupt the feature estimation, particularly if the least squares method is used. This problem is compounded in IR measurements because the underlying physical model describing the time-series is unknown, or is sometimes purposefully perturbed by the enemy.

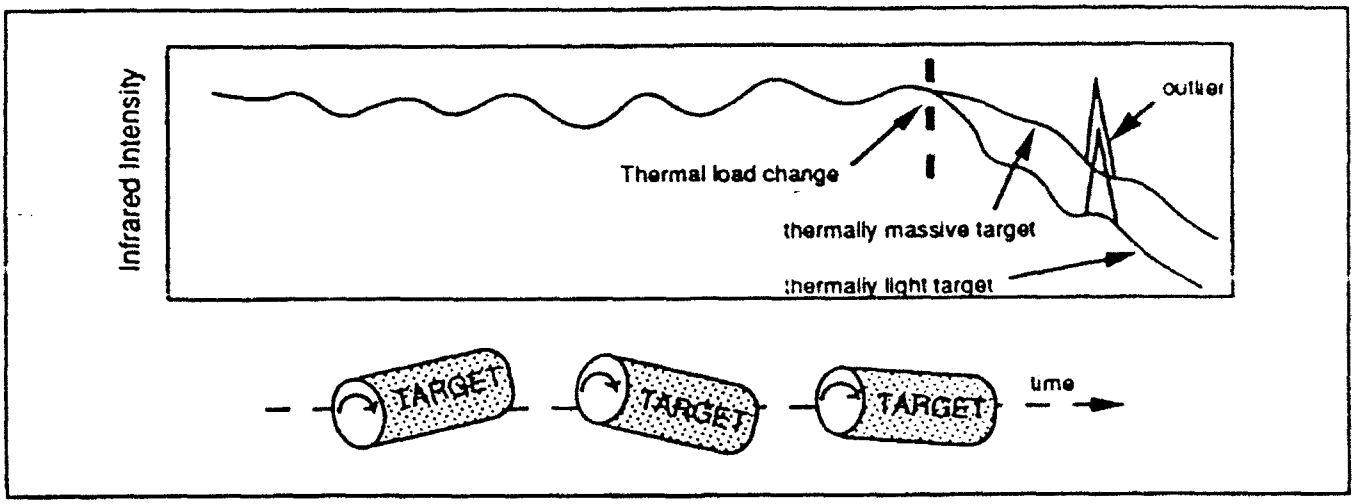

Figure 1) IR time-series is a sequence of IR measurements from a target that is moving, changing orientation and be subjected to a changing thermal load.

In a recent discrimination study for the Air Force Space Systems Division, we found that some preliminary results were misleading due to the presence of some specular data in our limited signature data base. Although an outlier removal routine had been applied to these signatures, not all of the specular data was removed. The remaining specular data had a significant impact on the preliminary performance results. J. Magnuson and M. Chernick devised an outlier removal routine based on the influence function for the variance of intensity. Unique features of the algorithm include the initial use (after detrending) of the middle 50% of the data and the adding back of observations with small to moderate influence. This avoids the masking problem. Empirically, the method appears to be effective since it removes the outliers without creating many "false alarms" (i.e., identifying valid observations as outliers). Preliminary work has been presented as a contributed paper at the annual meeting of the Institute of Mathematical Statistics in Washington D.C in August 1989 (Chernick and Magnuson (1989)).

This paper reports on the statistical properties of outlier detection algorithms based on the use of influence functions when specular behavior is present. We also consider one robust time series method due to Rousseeuw and Leroy (1987). Other approaches, such as Fox's test (see Fox (1972)) seem to suffer from the masking problem and are too sensitive to the detrending algorithm. We developed and tested these algorithms using simulated LWIR signatures. Currently we have a reasonably large data base of signatures for a variety of targets available. The methods used in Chernick and Magnuson (1989) are studied from the theoretical point of view. In particular, thresholds which were adjusted based on simulation results can be studied in the context of assumed probability models. Since observations are added back sequentially based on their ranking of intensity after detrending, recent developments in repeated significance testing and nonlinear renewal theory as developed by Woodroofe, Siegmund and Lai over the past 13 years are considered (see Siegmund (1985) and Woodroofe (1982) for details regarding this theory). Unfortunately due to the complicated nature of our algorithm, current theory only can be applied to partial sums in

one-parameter exponential families. It can therefore be applied to CUSUM tests but not to our sequential influence function algorithm.

Hampel's influence function was designed for observation vectors which are independent and identically distributed. The signature data we considered are correlated over time and are non-stationary. By removing the trends in the series we have removed the non-stationary behavior of the series. Trend removal is important and difficult when outliers are present. We compare various approaches. Proper trend removal is important since all the outlier detection algorithms depend on the time series being stationary. However the detrended series may still exhibit correlation and the methods applied to both the specular data and the closely spaced object (CSO) problems have thus far not taken this into account. In recent years Martin and his colleagues have looked carefully at the problems of parameter estimation and outlier detection for time series data. In particular they have defined influence functions which they believe are more appropriate for the detection of outliers in time series (Martin and Yohai (1986)). We have investigated the correlation structure of the detrended series. To date it has been difficult to find ways to make these generalizations practical. Lefrancois (1991) provides practical tests using empiric influence functions which we compare to our methods.

3. Editing Data with Closely Spaced Objects.

In strategic defense studies the surveillance sensors may not have the capability to resolve objects as early as would be desirable. Data from unresolved objects however may still be useful for tracking and discrimination. Depending on the sensor characteristics, the scenario for the deployment of targets and decoys and the sensor location, objects may, at various times, be unresolved. If signature intensities from two or more objects cannot be separated, the sensor will combine their intensities. If such data points are included in the signature, the track file can be so distorted as to severely degrade or destroy the discrimination performance. One remedy is to detect the times at which the track file is corrupted by multiple objects, to then remove these corrupted points (called CSOs for closely spaced objects) and discriminate on the basis of the remaining points. In recent work at Nichols Research Corporation, we have found that, for the scenarios we considered, some discrimination capability is still possible even though the simulated track files were often corrupted by CSOs. Our editing algorithms were based on the use of influence functions applied in an iterative fashion. Although the problem was similar to the specular data problem, there were some differences in the characteristics of the outliers (e.g. a much larger percentage of contamination for the CSO problem), the performance requirements and the necessary modifications used to get the algorithms to work properly. Again the statistical characteristics of the algorithms have not been studied. It would be a worthwhile research effort to look at the statistical properties of these algorithms when applied to signature data with closely spaced objects with an approach similar to our previous studies. Unfortunately, we did not have time to consider this in the current study.

4. Repeated Significance Tests

Our sequential outlier test will reject an observation as an outlier or specular occurrence if the observation, say X_i has a large influence on the variance of the detrended data which includes the i-1 observations from the middle of the detrended series and additionally those with the smallest influence on the variance. The decision criterion is, for fixed N (in our case N=61) to reject the first observation i as an outlier if

$$\frac{(X_i - \overline{X}_{(i-1)})^2}{S^2_{(i-1)}} - 1 > C,$$

where C is a given threshold, $\overline{X}_{(i-1)}$ is the mean of the middle 50% of the detrended observations plus any additional ones which have been added back and $S^2_{(i-1)}$ is the variance of the same i-1 observations.

This is a sequential hypothesis test where the test procedure is applied for several values of i. The threshold C can be constant or it can vary with N. If the test were to be applied for just one value i, it is usually straightforward to pick a value for C to control the type I and type II errors. However since the test is repeated for several values of i, the choice of C is more difficult.

The theory of repeated significance testing addresses this issue of choosing a threshold C. We hoped that the theory would be sufficiently general to be applied to our test. Unfortunately the theory is not sufficiently general and does not appear to be easy to extend to cover our problem. It may however still merit further study.

We briefly describe the theory of repeated significance tests and discuss the level of generality of the existing theory. For a more thorough description, the reader should consult Siegmund (1985) or Woodroofe (1982).

In the simplest setting we have n independent Gaussian random variables each with mean μ and variance σ^2 (denoted as $N(\mu,\sigma^2)$). Let $X_1, X_2, \ldots, X_n$ represent these n random variables. Define

$$S_n = \sum_{i=1}^{n} X_i .$$

Now $S_n/\sqrt{n}$ has the Gaussian distribution with mean $\sqrt{n}\,\mu$ and variance σ^2 (denoted $N(\sqrt{n}\,\mu,\sigma^2)$). Suppose that we want to test the hypothesis that $\mu=0$ versus the alternative $\mu \neq 0$.

From tables for the standard normal distribution (denoted N(0,1)) we have for fixed n and for σ^2 known

$$P[|S_n| > 3\sigma\sqrt{n}] = .0026 \text{ if } \mu=0,$$

since when $\mu = 0$, $S_n/\sqrt{n} \cong N(0,\sigma^2)$ or $S_n/(\sqrt{n}\sigma) \cong N(0,1)$.

Let the threshold C_n be $3\sigma\sqrt{n}$. If this threshold is exceeded for fixed n by $|S_n|$ then we have strong evidence for rejecting $\mu = 0$.

However, suppose we just continue to take observations until we find an n such that $|S_n| > 3\sigma\sqrt{n}$. This amounts to a sequential rule with stopping time

$$t = \inf\{n \geq 1 : |S_n| > 3\sigma\sqrt{n}\}.$$

However, from probability theory, the law of the iterated logarithm tells us that

$$\limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{2n \log\log n}} = 1 \text{ with probability } 1.$$

So even for $\mu = 0$

$\frac{S_n}{\sigma\sqrt{n}}$ will, with probability one, be of the order $\sqrt{2 \log\log n}$ for some n (actually for infinitely many values of n) which is certainly eventually bigger than 3. So the probability under the null hypothesis of ever finding such a t is 1.

Repeated significance testing amounts to choosing a maximum sample size N and rejecting the hypothesis that $\mu = 0$ if $|S_n| > c\,\sigma\sqrt{n}$ for some $n \leq N$. The probability of a false rejection $\alpha^* = P\{|S_n| > c\,\sigma\sqrt{n}, \text{ for some } n \leq N\}$. The choice of c is made to control α^*. The theory of repeated significance testing amounts to approximating α^* and related quantities for given threshold values c. The theory can be applied to one-parameter exponential families (a more general class of distributions than just the $N(\mu,\sigma^2)$ with σ^2 known)

With our influence function test, we have something like a repeated significance test. Things are much more complicated since we are dealing with a functional of the middle order statistics. However, since we add back observations and repeat testing the issue is the same. How do we determine a threshold which controls our probability of treating a "good" observation as a specular occurrence? Note also that the costs of the two error types are not necessarily the same. If we throw away a "good" observation, the cost is some loss of information regarding the feature estimates. On the other hand if we accept a "specular" occurrence, we are distorting the signature with potentially disastrous effects.

Although the determination of the threshold appears to be intractable with existing theory, bootstrap procedures as applied to hypothesis testing problems may provide a way to determine appropriate thresholds based on the data at hand and weak distributional assumptions. For a good up-to-date reference on bootstrap procedures in hypothesis testing problems see Fisher and Hall (1990).

5. Influence Functions

Hampel defined influence functions to measure the quality of robust estimators. Although the idea goes back to his 1968 Ph. D. dissertation, the most common reference in the literature is Hampel (1974). Implicit in the definition is the assumption that the observations are independent and identically distributed from a distribution F. In Hampel's definition the influence function depends on a point in the observation space usually taken to be a particular observation, the parameter being estimated expressed as a functional of the distribution and the distribution F itself. The general theory as applied to robust statistics is covered in Hampel et al. (1986).

For example, the variance σ^2 of a uni-variate distribution F can be expressed as a functional of F denoted by T(F), where μ is the mean of F and

$$T(F) = \int_{-}^{-} (x-\mu)^2 \, dF .$$

In general the influence function is defined as a type of directional derivative of the functional in the space of distributions.

Let I(F,T(F),x) denote the influence function at the point x for the functional T(F) with distribution F. Definition:

$$I(F,T(F),x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon}$$

where ε is a positive real number and δ_x is the distribution with all its probability concentrated at x. The distribution $G = (1 - \varepsilon)F + \varepsilon\delta_x$ is a mixture distribution which is "close" to F for small ε. This definition is quite general and applies to bivariate and multi-variate distributions F as well as the uni-variate distribution we are considering in this discussion.

Influence functions are similar to derivatives of functions and retain many properties from calculus, including the product rule, the quotient rule and the chain rule.

From the above definition, we shall derive the influence function for the variance. Many other influence functions have been given in the literature. Gnanadesikan (1977) gives the influence function for bivariate correlation which he attributes to Mallows in an unpublished paper. See Chernick and Murthy (1983) for some other simple examples.

Now we consider I(F,T(F),x) for

$$T(F) = \int_{-}^{-} (y-\mu)^2 \, dF(y)$$

$$I(F,T(F),x) = \lim_{\varepsilon \to 0} \frac{T(G) - T(F)}{\varepsilon} \qquad \text{where}$$

$$T(G) = \int_{-}^{-} (y-m)^2 \, dG(y) \qquad \text{and} \qquad m = \int_{-}^{-} y \, dG(y)$$

$$\text{First } m = \int_{-\infty}^{\infty} y \, dG(y) = \int_{-\infty}^{\infty} (1-\varepsilon)\, y \, dF(y) + \varepsilon x = (1-\varepsilon)\mu + \varepsilon x$$

So $$T(G) = \int_{-\infty}^{\infty} (y - (1 - \varepsilon)\mu - \varepsilon x)^2 \, dG(y) = \int_{-\infty}^{\infty} (y - \mu - \varepsilon(x - \mu))^2 \, dG(y)$$

$$= (1-\varepsilon) \int_{-\infty}^{\infty} \left[(y - \mu)^2 - 2\varepsilon(y - \mu)(x - \mu) + \varepsilon^2(x - \mu)^2\right] dF(y) + \varepsilon\left[(x - \mu)^2 - 2\varepsilon(x - \mu)^2 + \varepsilon^2(x - \mu)^2\right]$$

So $T(G) - T(F) =$

$$\varepsilon(x - \mu)^2 (1-2\varepsilon+\varepsilon^2) - \varepsilon \int_{-\infty}^{\infty} (y - \mu)^2 \, dF(y) + (1-\varepsilon) \int_{-\infty}^{\infty} \left[\varepsilon^2(x - \mu)^2 - 2\varepsilon(x-\mu)(y-\mu)\right] dF(y)$$

Since $\int_{-\infty}^{\infty} (y - \mu) \, dF(y) = 0$, we simplify to obtain

$$T(G) - T(F) = \varepsilon(x-\mu)^2(1-2\varepsilon+\varepsilon^2) - \varepsilon \sigma^2 + (1-\varepsilon)\varepsilon^2(x-\mu)^2,$$

Noting that $\sigma^2 = \int_{-\infty}^{\infty} (y - \mu)^2 \, dF(y)$ and $\int_{-\infty}^{\infty} dF(y) = 1$

Dividing by ε yields

$$\frac{T(G) - T(F)}{\varepsilon} = (x-\mu)^2(1-2\varepsilon+\varepsilon^2) - \sigma^2 + \varepsilon(1-\varepsilon)(x-\mu)^2$$

$$= (x-\mu)^2(1-\varepsilon) - \sigma^2$$

Now taking the limit as ε approaches 0 we have

$$\lim_{\varepsilon \to 0} \frac{T(G) - T(F)}{\varepsilon} = (x-\mu)^2 - \sigma^2 \qquad (1)$$

This shows that the influence function for the variance depends on x but only depends on F through μ and σ^2. The derivation applies to any distribution F with finite second moments.

Simple formulas such as (1) above are useful in approximating the influence function when the parameters μ and σ^2 are unknown. We simply replace them with their sample estimates. Hence an empirical estimate of the influence function for the variance at X would be

$$\hat{I} = (X - \bar{X})^2 - S^2$$

where $\bar{X}$ is the sample mean and S^2 is the sample variance.

The influence function has a very useful interpretation. Suppose that we replace X with the observation X_i and replace F by F_{n-1} the empirical distribution (i.e. the distribution with probability 1/(n-1) on each observation). For large n, F_{n-1} approximates F and we can take $\varepsilon = 1/(n-1)$ since 1/(n-1) is small. We see then that

$$T(G) = T(\frac{n-1}{n} F_{n-1} + \frac{1}{n} \delta x) = T(F_n)$$

So

$$\frac{T(G) - T(F_{n-1})}{\varepsilon} = \frac{T(F_n) - T(F_{n-1})}{\left(\frac{1}{n-1}\right)}$$

$$= (n-1)(T(F_n)-T(F_{n-1}))$$

Now $T(F_n)$ *is the sample variance when* X_i *is included and* $T(F_{n-1})$ *is* the sample variance when X_i is excluded. So the influence function approximates the difference between the estimate with the observation included and with the observation excluded multiplied by the sample size. With this interpretation we see that the influence function is a very appealing sensitivity measure of the effect of an observation on the estimate. Such measures are useful in determining outliers with respect to particular parameters of interest (in this case the variance).

Strictly speaking these results apply only to independent and identically distributed observations, but we have found them to be useful in practice even when the observations are correlated (e.g. Chernick et al. (1982) and Chernick and Magnuson (1989)). Martin and Yohai (1986) have generalized Hampel's definition to time series data and Lefrancois (1991) uses some practical and easily computable forms of influence functions which have a similar appealing notion of sensitivity of the estimates to the observations.

The sequential influence function outlier detection algorithm simply divides by σ^2 to remove any scale dependence in setting a threshold. So instead of using $(X_i - \bar{X})^2 - S^2$ and comparing it to a threshold C, we use

$$\frac{(X_i - \bar{X})^2}{S^2} - 1$$

and compare it to a threshold C. In our case, the detrending algorithm is applied first and the middle 50% of the detrended observations are used to initially estimate μ and σ^2. This is done to avoid masking due to estimate sensitivity to outliers.

A detailed description of the outlier detection algorithms used to detect the specular occurrences is given in the next section.

6. Algorithms Studied

In our investigation we considered four approaches to the outlier detection problem. Fox (1972) provided one of the earliest approaches to identifying outliers in time series. His was a pioneering paper which is often cited in the literature. Fox's major contribution was to identify two types of outliers which are referred to as (1) additive outliers and (2) innovation outliers. A model which considers only additive outliers is referred to as an AO model and one which considers only innovation outliers is referred to as an IO model. These models have been adopted by Martin and others in much of the subsequent work on robust time series modeling and outlier detection.

The distinction made between AO and IO outliers is that the AO outlier is a unique occurrence which has an additive effect on a single observation but does not effect subsequent observations, whereas the IO outlier affects a particular observation and all subsequent observations. The specular occurrences should be modelled as AO outliers since they have short duration and do not affect the magnitude of subsequent observations. On occasion a specular occurrence may last for more than one time measurement interval. In such cases it may be best to model the event as possible multiple AO outliers.

Fox's other contribution was to provide likelihood ratio tests for detecting an AO or an IO outlier in a time series. Although this work has value, it is somewhat restrictive. First it requires that the time series is stationary and well approximated by a low order autoregressive process. Fox states in the introduction "Throughout this paper, trend and seasonal components are assumed either negligible or to have been eliminated. The method adopted to remove these components might affect the results in some way." In fact our signatures have significant polynomial trends and periodic components which need to be removed first. Since the presence of the outliers can have an impact on the success of the detrending algorithm it also has an affect on the ability of Fox's algorithm to detect the outliers. This difficulty, which is more severe for some of the algorithms, is a problem which is faced by all algorithms.

Fox's tests are designed to have maximum power (i. e. highest probability of detecting an outlier) when the entire series has a single outlier (either AO or IO) and the location of the outliers is known. It also can have practical utility when the locations of the outliers are unknown. Unfortunately, its biggest weakness is that it is not designed for dealing with multiple outliers. Our preliminary investigations showed that masking was a serious problem with Fox's test. Masking occurs when the presence of multiple outliers inhibits parameter estimation which in turn reduces the ability to differentiate outliers from "good" data. For signatures with specular occurrences, multiple occurrences are quite common. For this reason, Fox's test is not competitive with our sequential influence function test and thus was dropped from the comparison.

Chernick et al. (1982) developed an influence function matrix for the autocorrelation function of a stationary time series. The influence function for the correlation at lag k was determined by analogy to the bivariate correlation of the components of independent random two - dimensional vectors as given in Gnanadesikan (1977). They showed through simulations and some real data examples the practical value of this influence function matrix for detecting multiple outliers in the time series data. Later Martin and Yohai (1986) devised more general influence functions which are appropriate for correlated data. Hampel's influence function really is meant to apply only to independent identically distributed random variables or vectors.

Although Martin and Yohai (1986) provide a suitable general theory, it is difficult to infer a practical influence function measure from their results. Lefrancois (1991) has provided practical measures of influence for time series. He provides thresholds for these influence functions based on the assumption that the time series is stationary and Gaussian. In referring to earlier work on influence functions for time series, Lefrancois states that "almost all were developed without recourse to general appropriate theory, and without critical values for declaring an observation as over - influential or not. Only Chernick et al. (1982) mentioned a threshold, but it does not correspond to the distribution theory of their suggested measure."

Actually Chernick et al. (1982) did derive thresholds based on the product standard normal distribution which is appropriate when one is willing to assume the process is Gaussian. They also derived the distribution for a summary test statistic which they called the average squared influence function. This statistic has asymptotically a chi-square distribution. They were, however, reluctant to apply the distribution theory to the real data examples because they felt that the Gaussian assumptions were not justifiable.

It is important to note that the formulas for influence functions only require the existence of certain moments. In the case of autocorrelations, we need second moments. Large values of the influence function will indicate outliers even for many non-Gaussian time series. Chernick et al. (1982) chose to use thresholds based on empirical experimentation with the observed data. At this time bootstrap procedures could be used to arrive at thresholds based on the observed time series, avoiding Gaussian assumptions.

We believe that the measures provided in Lefrancois (1991) are appropriate and useful for detecting outliers in time series. The thresholds he derives are appropriate for Gaussian processes but may not be appropriate otherwise.

Although Lefrancois states "We consider only the case of at most one over-influential observation.", he does demonstrate its use in paragraph 4 as a method for detecting outliers (i.e. multiple outliers) in time series. Although influential observations need not be erroneous or outliers, in practice they often are and we believe Lefrancois' remarks are overly cautious.

We consider in this paper Lefrancois' sample influence measure which he denotes $SIC_{i,k}$. Let μ and σ^2 denote the mean and the variance of a stationary time series Xi i=1,2,...,n. We define

$Z_i = (x-\mu)/\sigma$. Actually in practice, since μ and σ^2 will be unknown, sample estimates will be used in their place in computing Z_i.

Lefrancois' $SIC_{i,k}$, is the sample influence of observation i on the lag k autocorrelation r_k. From equation (2.4) in Lefrancois (1991) we approximate $SIC_{i,k}$ by

$$SIC_{i,k} = (Z_i Z_{i+k} + Z_i Z_{i-k} - r_k Z_i^2) / \{1 - Z_i^2 / (n-1)\}$$

This is contrasted with the Chernick et al. (1982) measure which is

$$C_{i,k} = Z_i Z_{i+k} - \frac{1}{2}(Z_i^2 + Z_{i+k}^2) r_k$$

Note that for large n the denominator in $SIC_{i,k}$ is close to 1. Ignoring the denominator we see a great deal of similarities between the two equations above. $-1/2(Z_i^2+Z_{i+k}^2)r_k$ in $C_{i,k}$ is about the same as $-Z_i^2 r_k$ in $SIC_{i,k}$ and the term Z_iZ_{i+k} occurs in both equations. The main difference is the term Z_iZ_{i-k} in $SIC_{i,k}$ which does not appear in $C_{i,k}$. Lefrancois points to this missing term in Chernick et al. (1982). We believe that this point and the 'leave-one-out' interpretation of $SIC_{i,k}$ justifies the belief that it is an improvement over Chernick et al. (1982). Lefrancois (1991) computes a summary statistic QIC_i which is a quadratic form obtained from $SIC_{i,k}$. This is similar to the average squared influence function of Chernick et al. (1982). To test for a single highly influential observation, he obtains Bonferroni type upper and lower bounds on the probability distribution for the largest QIC_i in the time series. To detect multiple outliers this test can be applied sequentially (i.e. test the largest and if it exceeds the threshold remove it and repeat the procedure on the time series with that observation left out). The sequential procedure is continued until no outliers remain (i.e. the largest QIC_i no longer exceeds the threshold). This procedure appears to have worked well in the example of paragraph 4 in Lefrancois (1991) but masking may still be a problem with this approach.

In section 5, we derived Hampel's influence function at x for the variance of independent identically distributed observations from a distribution with finite second moment. For a stationary time series we interpret this result as the effect an observation x will have on the variance of the time series (although strictly speaking the result does not apply to correlated observations). Since specular occurrences have a large effect on the variance of the signature, we expect that such an influence measure would be able to detect these occurrences. This was borne out in Chernick and Magnuson (1989) and has been used at NRC to edit simulated signatures since that time.

As we pointed out in Section 4, repeated significance testing provides the appropriate framework for determining a threshold for our test statistic. Since the results do not appear tractable, empirically defined thresholds were used. A better approach using bootstrap hypothesis tests may be the subject of further research. Our test is essentially to reject for large values of Z_i^2 where $Z_i=(x_i-\mu)/\sigma$ as before. However since m and s are unknown and estimates are sensitive to outliers, they are replaced by $\overline{X}_{(i-1)}$and $S^2_{(i-1)}$ as defined earlier. This is very important since it essentially removes the masking problem by truncating the data and then adding back observations until the outliers are found. A similar approach to Lefrancois' statistics might also help his algorithm even though it makes the distribution theory intractable. Both our sequential influence function for variance algorithm and Lefrancois' influence function for autocorrelation are based on the fact that least squares estimation is sensitive to outliers. So outliers or specular occurrences will cause large differences in estimates such as variance and autocorrelation. Measures which can estimate this effect can therefore detect the specular occurrences.

Our naive intuition tells us that if we have an appropriate model for the "good" data, then the outliers should be detected because they have a large residual (i. e. deviation of the observed value from what the model expects). With least squares estimation, however these outliers can have such a high influence on the model parameters that their residuals are small! An alternative approach is to find estimation procedures which are insensitive to the outliers (i.e. robust time series modelling) and then detect the outliers based on residuals from the "robust" model. In recent years, motivated by results in robust regression modeling, there have been a number of procedures devised to obtain robust estimates of the parameters in a time series model. In order to detect multiple outliers, we seek a procedure

which doesn't breakdown until many outliers are present. In robust regression the repeated median algorithm of Siegel (1982) is an example of a procedure with an asymptotic breakdown point when 50% of the data are outliers. It achieves the maximum possible breakdown value in the limit as the sample size gets larger and larger.

Of the available methods in the time series literature, the most promising appears to be the least median of squares algorithm of Leroy and Rousseeuw. The procedure is to fit robustly a low order autoregressive model to the stationary time series. Based on fitting some detrended signatures without speculars, we decided to apply an autoregressive model of the fourth order (AR(4)) to the data. It appears that it may be better to remove the remaining seasonal components first.

The least median of squares algorithm considers the square residuals just as does least squares but instead of minimizing the sum of the squared residuals, it minimizes the median of the entire set of squared residuals. The outliers can then produce large squared residuals and have no effect on the estimate since their squared residuals are well above the median. The procedure is conceptually simple but difficult to compute. The basic procedure is described by Rousseeuw and Leroy (1987) pp 197-204. It is very computer intensive and tricks are provided to reduce the computation. For time series models a slight modification is required. This is described on page 279 of Rousseeuw and Leroy (1987). An example of the successful application of the procedure is given on pages 279-284. After fitting the model, the observations with large absolute residual are considered to be outliers. In the case of specular occurrences, this test could be one-sided (i.e. only observations which are much larger than the model prediction will be considered).

7. Results of Algorithm Comparisons

7.1) Detrending algorithms

The minimum window performed the best since it takes advantage of our apriori knowledge that outliers are always large. The median window detrending algorithm performance was worse for our special case, but it is more general since it makes no assumptions about the direction of the outliers. The median window detrending algorithm is recommended for general outlier detection. The Fourier series detrending algorithm was severely affected by outliers. Its performance was inferior.

7.2) Equal error rate as a fair scoring parameter

Since we are comparing the results of several combinations of different algorithms we need a common scoring technique. The equal error rate was chosen as a fair and consistent scoring method for several reasons. Figure 2 is a schematic of how the equal error rate is calculated. The histograms on the left in Figure 2 are the distributions of good data and outliers. The outlier distribution tends to lie to the right of the good data because the outliers tend to have higher intensity than the good data. The threshold figure 2a is placed in an arbitrary position for purposes of explanation. Any data lying to the left of the threshold is designated good data and everything to the right is designated as an outlier. A type 1 error, an outlier mistakenly designated as good data, occurs for those outliers in the tail of the distribution which lie to the left of the threshold. A type 2 error, good data mistakenly labeled as an outlier, occurs for the good data in the tail of the distribution which lie to the right of the threshold. Figure 2b is calculated by moving the threshold through a range of values. If the threshold is arbitrarily placed to the left of both distributions, all the data will be assumed to be outliers and 100% of the good data will be discarded (type 2 error). As the threshold is moved to the right, less of the good data will be

discarded but more outliers will be retained, type 1 errors increase while type 2 errors decrease.

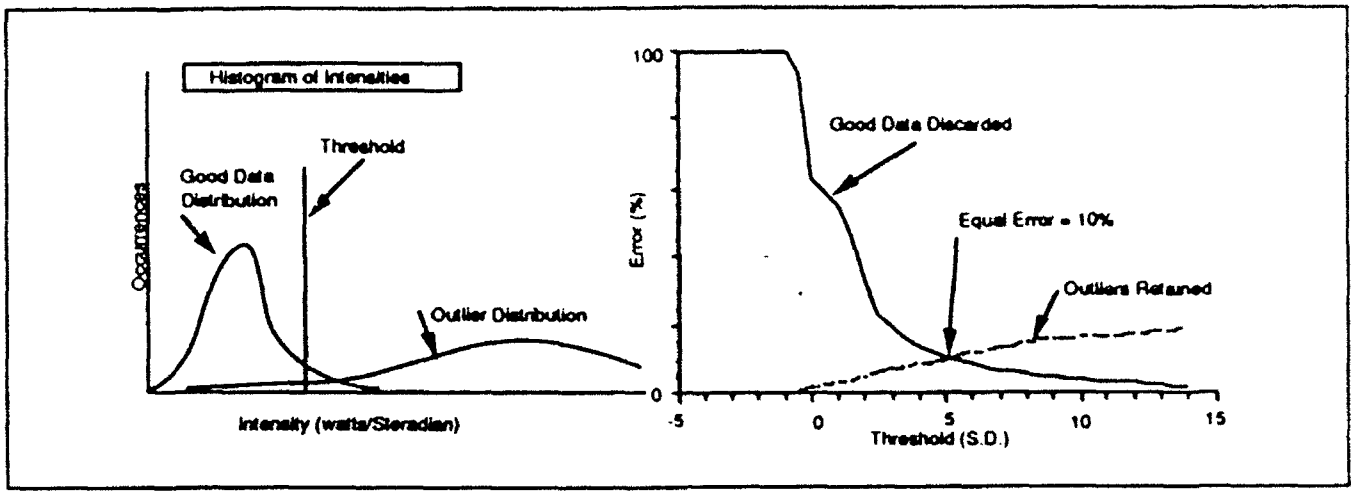

Figure 2) Calculation of equal error rate.

The exact placement of the threshold depends on how well the two different types of errors can be tolerated. If outliers are extremely detrimental, the threshold should be moved to the left as far as possible. On the other hand, if good data must be retained at all costs, then the threshold should be moved to the right as far as possible. No matter where the threshold is fixed, there is always a compromise between the two types of errors. A fair compromise is to place the threshold where the two types of errors are equal, thereby the name "equal error rate". This is a traditional placement when the detrimental effects of the different types of errors are not known in advance.

7.3) Fox's algorithm

Since Fox's algorithm is not intended for multiple outliers, its performance was expected to be poor due to the masking problem. The results were 43% equal error for the minimum window detrending algorithm, 46% equal error for the median window detrending algorithm and 49% equal error for the Fourier filter detrending algorithm.

7.4) Sequential Influence Function

The performance for the Sequential Influence function was much better. The results were 12% equal error for the minimum window detrending algorithm, 17% equal error for the median window detrending algorithm and 30% equal error for the Fourier filter detrending algorithm. The main improvement was accomplished because of overcoming the masking problem.

7.5) Sequential AR(1) Influence Function

A slight, but significant improvement resulted when a first order autoregressive (AR) model was used. The results were a 9% equal error for the minimum window detrending algorithm. The characteristic operating curve is plotted in Figure 3.

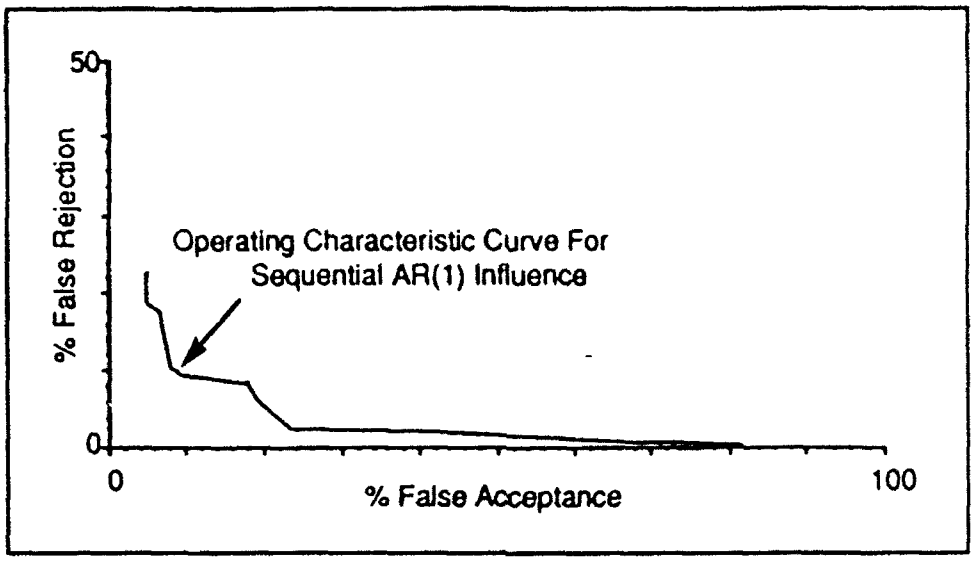

Figure 3) The characteristic operating curve for the Sequential AR(1) influence function algorithm.

7.6) Lefrancois' algorithm

Lefrancois' algorithm requires an intermediate step of calculating an influence function SIC. An overly influential point in the time series will not only perturb the influence function for the time point, it will also perturb the influence function of adjacent points at different lags in what has been coined the "*clothes-pin*" *effect*. The clothes pin shape is displayed in Table 1. The clothes-pin effect stands out when there is only one outlier in a time series. Even with only one outlier, however there are some coincidences where a relatively small data point, like number 54 which is not an outlier, can serendipitously have a large sum of squares and become suspicious.

Unfortunately, when several outliers are present, the estimates of the autocorrelation parameters are affected and the influence function are badly corrupted as in Table 2.

The value of the influence function for point number 31 in Table 2 is relatively large, 1.42, compared to the values for the two outliers at points number 23 and 27. There were other outliers in this series which also corrupted the autoregressive parameter estimation.

Influence functions are also affected by large values in data which really have no outliers. Table 3 is an example of such a series. This series has a large oscillatory component which may or may not be adequately modeled by the fourth order autoregressive process AR(4).

The overall performance for data with an arbitrary number of outliers is described in Figure 4. The equal error operating point is 31% false acceptance and false rejection. This is considerably worse than the 9% equal error rate established by the sequential influence function. The major difference is that in the Lefrancois algorithm, the autoregressive parameters were calculated using all the data and then the points with the largest influence functions were removed. The opposite approach was taken with the sequential influence function.

Obs Num	Data	Lag				Sum^2/N
		1	2	3	4	
44	0.8625	-0.08	-0.18	-0.01	-0.17	0.11
45	1.1259	-0.44	-0.73	0.27	-0.20	0.16
46	0.0000	-0.09	-0.14	1.33	0.73	0.65
47	0.4048	0.27	0.14	0.22	**-1.07**	0.33
48	0.9925	-0.55	-0.90	**-2.00**	0.04	0.66
49	0.1732	0.56	**-2.10**	0.62	0.21	0.91
50	0.0000	**-2.77**	0.36	0.95	1.01	1.42
51	2.9560 ***	**-2.41**	**1.87**	**-6.70**	**-2.13**	9.50
52	0.5609	-0.43	0.20	0.07	0.09	0.82
53	0.0000	0.40	-3.42	1.62	1.30	2.60
54	0.6798	-0.01	-0.01	0.03	-0.01	5.77
55	0.8580	-0.29	-0.41	-0.14	0.75	2.14
56	0.0000	0.48	0.84	1.51	1.21	0.93
57	0.3332	0.69	0.53	0.40	0.77	0.45
58	0.5865	0.20	0.30	0.01		0.19
59	0.0665	1.32	1.51			1.20
60	0.0000	1.71				1.27
61	0.3876					1.46

Table 1) Lefrancois Influence Functions for data with only one outlier at time point 51. The "clothes-pin" effect is emphasized by the numbers which are bold and underlined SIC values. The outlier is marked by ***.

Obs Num	Data	Lag				Sum²/N
		1	2	3	4	
19	0.0302	0.32	0.30	-0.51	-0.45	0.24
20	0.0459	0.30	0.30	-0.41	-0.51	0.32
21	0.0051	0.35	-0.51	0.35	0.35	0.13
22	0.0000	-0.48	0.31	0.37	0.35	0.15
23	1.2388 ***	-0.90	-1.38	-0.83	2.58	1.37
24	0.0402	-0.41	0.29	-0.49	0.28	0.23
25	0.0000	0.34	-1.46	0.39	0.37	0.61
26	0.0340	-0.50	0.29	-0.42	0.30	0.20
27	1.3696 ***	-0.97	-1.59	-0.89	2.57	2.52
28	0.0049	-0.56	0.31	0.36	0.34	0.24
29	0.0000	0.36	-0.62	0.37	-1.32	0.69
30	0.0428	0.31	0.30	-1.90	0.27	0.62
31	0.0291	0.32	-1.18	0.31	-1.59	1.42
32	0.0000	-1.33	0.31	-0.81	0.37	0.39

Table 2) *Lefrancois Influence Functions* with multiple outliers

Obs Num	Data	Lag				Sum²/N
		1	2	3	4	
27	1.1609	-0.35	1.10	-0.83	3.33	8.49
28	0.4309	-0.02	0.02	-0.17	-0.14	0.24
29	0.0000	0.08	-1.26	-1.79	-1.02	1.00
30	0.5672	-0.10	0.39	0.05	0.30	0.12
31	0.5351	0.29	-0.16	0.24	0.07	1.62
32	2.1514	-0.02	3.68	-4.80	6.72	10.65
33	0.0000	-3.90	-0.01	-1.24	0.71	2.29
34	1.9764	-1.83	7.69	-0.09	0.40	11.43
35	0.1699	-1.84	0.20	-1.00	-0.12	3.85
36	1.3120	-0.08	1.36	-0.86	4.67	16.70
37	0.5404	0.07	-0.03	0.28	-0.16	0.08
38	0.0000	-0.17	-1.49	0.93	-2.54	1.57

Table 3) Lefrancois Influence Functions - data with no outliers

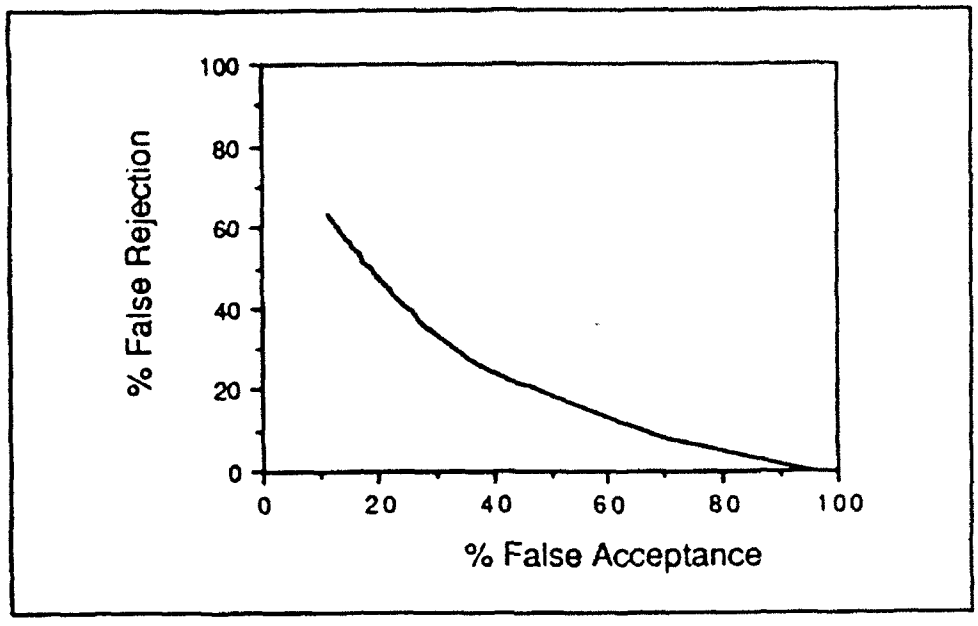

Figure 4) Operating Characteristic Curve for Lefrancois Influence Function Outlier Detection Algorithm.

8. Human Experiment

In fact humans seem to be excellent outlier detectors, out performing all the algorithms tested so far. Relles and Rogers (1977) showed that statisticians can subjectively estimate a location parameter in the face of outliers quite well. Similarly we demonstrate that for time series signatures, engineers are good at identifying outliers. As an experiment, 4 volunteers were tested as outlier detectors. They were given 30 Infra-red measurement time-series with 61 time points in each series. Appendix A contains four of the 30 samples. The volunteers were told that there were outliers in some, but not all of the series and asked to circle all points that they considered to be outliers. All four people have engineering degrees. They had varying degrees of familiarity with time series. Figure 5 shows a comparison between the best computer algorithm and the volunteers. Not surprisingly, the engineer with the most experience with time series performed the best, he correctly identified all but 2.6% of the outliers and erroneously mislabeled only 0.5 % of the "good" data as outliers. The next most experienced engineer performed second best. She missed 16.7% of the outliers but did not make any mistakes of throwing away good data. Figure 5 is a comparison between the operating characteristic curve of the best algorithm and the volunteers. The operating characteristic curve is derived by varying the threshold on the outlier detection algorithm. There is a trade-off between threshold value and the performance of the algorithm. On the one hand, if the threshold is lowered, more "good" data will be falsely rejected and fewer outliers will be falsely accepted. On the other hand, if the threshold is raised, less "good" data will be falsely rejected, but more outliers will be falsely accepted.

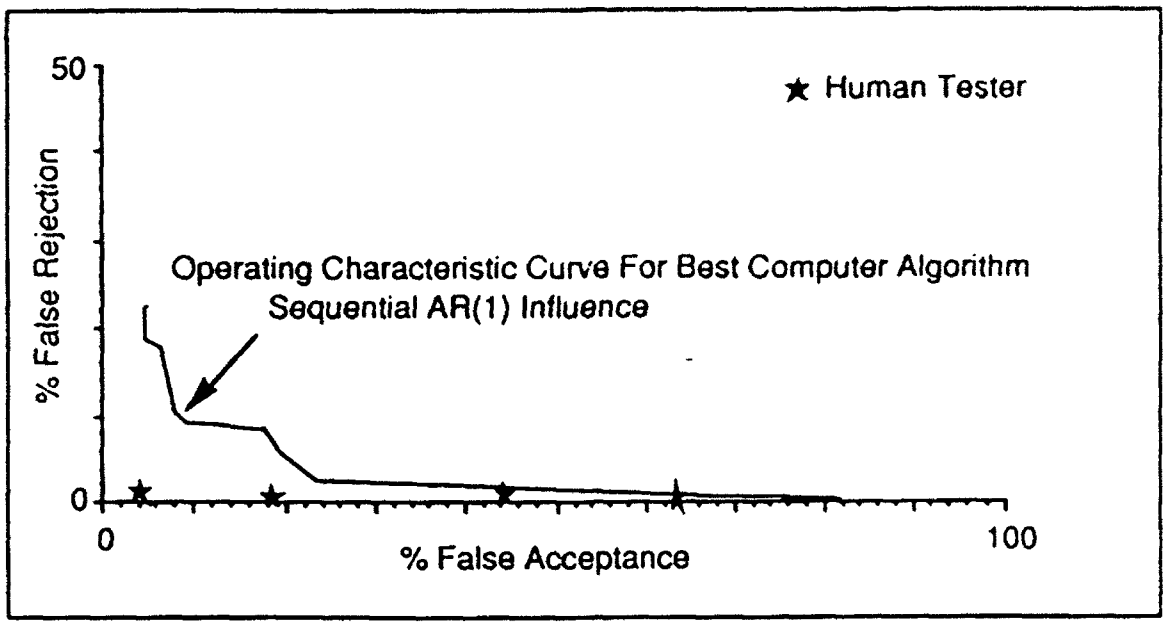

Figure 5) Comparing human outlier detectors and the best computers algorithm. There is a trade off between false acceptance of outliers and false rejection of "good" data. The volunteers who had a high acceptance rate were conservative in that they did not want to throw away any "good" data.

Four time-series plots are included in Appendix A. These four plots are the scored results from the best human outlier detector, they help explain the problems associated with outlier detection.

The first plot, figure ZX-140.DAT, shows the problem of detecting outliers when the time series is non-stationary. The tester correctly identified all three outliers in this plot with ease, even though one of the outliers lies on a steep slope. Computer detrending algorithms are readily available for data with no outliers, but they are badly corrupted when multiple outliers are present. This is a chicken and egg problem. If the data were detrended, outlier detection would be simple, however, the outlier's presence corrupts the detrending algorithm to the point where it is sometimes difficult to detect it.

The second plot, figure ZX-141.DAT, shows the first step in the human's outlier detection algorithm. He drew an envelope around the range of "good" data. This is somewhat similar to whitening and detrending with a variable variance filter. The outliers are readily identifiable outside the expected envelope.

The third plot, figure ZX-131.DAT shows the second improvement from the human detector. Once the large outliers were removed, he established the rough period of oscillation in the time-series and looked for patterns where the points were perturbed from the sinusoid shape. Amazingly he only missed 1 outlier in the whole series.

Plot four, figure ZX-151.DAT almost tricked the best volunteer. Even though there are no outliers in this time-series, the modulation between the sampling and the oscillations in the IR object produced large spikes which at first look like outliers. The volunteer reconsidered them and rightfully said that "all OK, no outliers". This shows the problem with outlier detection, a variable adaptive detrending algorithm would be fooled by the last signature.

9. Conclusions

In conclusion, the Sequential AR(1) Influence Function preformed better than any other computer algorithm on the data tested in this study, see table 4.. The algorithm works well, in part, because the sequential nature of the algorithm avoids the masking problem that happens when several outliers are present in the data.

Fox's Algorithm	43.0%
Lefrancois Influence Function	31.0%
Sequential Influence Function	12.0%
Sequential AR(1) Influence Function	9.0%
Human volunteers	2.6%

Table 4) Summary of equal error, which is a measure of the ability of each algorithm to discriminate outliers from good data.

Due to massive data processing requirements and the need for real time outlier detection, it would be impractical to use humans as outlier detectors, even though they easily outperform all the computer algorithms. The human volunteer test was informative for several reasons. It shows how well a computer algorithm could be expected to perform if it could capture the expertise of a human. The study also helps illustrate two major problems with IR signatures..

1) Humans can handle changes of slope where detrending algorithms have difficulty. Many good detrending algorithms are available for data without corruption by outliers. None that we know of work well when there are several outliers present. This is a problem that should be looked into in more detail.

2) Several outliers mask each other so estimating influence function parameters can be badly corrupted by the outliers. A better approach is to establish robust estimates of the parameters and then test for influence.

Because the enemy has control of the IR signature, the underlying model of the target IR time series should be established independent of outliers. There is great hope for overcoming most of these problems and establishing an outlier detection algorithm which is nearly as good as the human expert once a good outlier resistant detrending algorithm is developed.

References

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control, Second Edition, Holden-Day, San Francisco.

Barnett, V. and Lewis, T. (1984) Outliers in Statistical Data, Second Edition, John Wiley and Sons Inc., New York.

Chang, I., Tiao, G. C. and Chen, C. (1988). "Estimation of Time Series Parameters in the Presence of Outliers.", Technometrics 30 193-204.

Chatterjee, S. and Hadi, A. S. (1988), Sensitivity Analysis in Linear Regression, John Wiley and Sons Inc, New York.

Chernick, M. R. (1980) "A Note on the Robustness of Dixon's Ratio Test in Small Samples," Oak Ridge National Lab/TM-7625.

Chernick, M. R. (1982a) "The Influence Function and Its Application to Data Validation," American Journal of Mathematical and Management Sciences, 2, 263-288.

Chernick, M. R. (1982b), "A Note on the Robustness of Dixon's Ratio Test in Small Samples," The American Statistician, 36, 140.

Chernick, M. R. Downing, D. J. and Pike, D. H. (1982), "Detecting Outliers in Time Series Data", Journal of the American Statistical Association, 77, 743-747.

Chernick, M. R. and Magnuson, J. A. (1989), "An Algorithm to Detect Specular Occurrences in Infrared Sensor Signatures Based on Influence Functions," Presentation at the Institute of Mathematical Statistics Annual Meeting in Washington D. C..

Chernick, M. R. and Murthy, V. K. (1983), "The Use of Influence Functions for Outlier Detection and Data Editing," American Journal of Mathematical and Management Sciences, 3, 47-61.

Fisher, N. I. and Hall, P. (1990). "On Bootstrap Hypothesis Testing," Australian Journal of Statistics 32, 177-190.

Fox, A. J. (1972). "Outliers in Time Series", Journal of the Royal Statistical Society Series B, 43, 350-363.

Gnanadesikan, R. (1977), Methods for Statistical Data Analysis of Multivariate Observations, John Wiley and Sons Inc., New York.

Hampel, F. R. (1974). "The Influence Curve and its Role in Robust Estimation," Journal of American Statistical Association. 69, 383-393.

Hampel, F, R. Rousseeuw, P. J, Ronchetti, E. M. and Stahel, W. A. (1986). Robust Statistics: The Approach Based on Influence Functions, John Wiley and Sons Inc. New York.

Huber, P. J. (1981). Robust Statistics, John Wiley and Sons Inc., New York.

Lattin, J. M. (1983). "Identifying Influential Observations in Time Series Data." In Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface, Ed. K. W. Heiner, R. S. Sacher, and J. W. Wilkinson, 283-286, Springer-Verlag, Berlin.

Lefrancois, B. (1991). "Detecting Over-influential Observations in Time Series" Biometrika 78, 91-99.

Li. W. K. and Hui, Y. V. (1987). "On the Empirical Influence Functions of Residual Autocorrelation in Time Series . In Proceedings of Business Economic Statistics Section, American Statistical Association, 465-468.

Martin, R. D. and Yohai, V. J. (1986), "Influence Functions for Time Series," The Annals of Statistics, 14, 781-855.

Naus, J. I (1975), Data Quality Control and Editing, Marcel Dekker Inc, New York.

Relles, D. A. and Rogers, W. H. (1977). "Statisticians are Fairly Robust Estimators of Location." Journal of the American Statistical Association 72, 107-111.

Rousseeuw, P. J. and Leroy, A. M. (1987) Robust Regression and Outlier Detection, John Wiley and Sons Inc., New York.

Siegel, A. F. (1982). "Robust Regression Using Repeated Medians", Biometrika 69, 242-244.

Siegmund, D. (1985), Sequential Analysis: Tests and Confidence Intervals, Springer-Verlag, New York.

Walker, E. (1989) "Detection of Collinearity Influential Observations," Communications in Statistics, Theory and Methods, 18, 1675-1690.

Woodroofe, M. (1982), Nonlinear Renewal Theory in Sequential Analysis, Society for Industrial and Applied Mathematics, Philadelphia.

Appendix A - Four sample IR signatures used to test human ability to detect outliers.